- You must show steps with proper justification in your solutions. Partial credits are given to intermediate steps and reasonings. At the same time please keep your answer concise with important key points/steps.
- Please submit only pdf or image files (either PNG or JPG/JPEG). If you submit OS specific files (such as .pages files) and/or corrupted files, your assignment will not be graded.
- Use R to plot graphs and obtain summary statistics
- You must attach all plots within the question when you hand in the assignment and ensure that all your graphs are labelled appropriately. (i.e. do not put all the graphs at the end of the assignment)
- Answers must be written in the context of the question where applicable.

1. The state of New York has one of the heaviest air traffic in the world. In 2013, over 300,000 flights left from or arrived at the biggest three New York airports: La Guardia Airport (LGA), John F. Kennedy International Airport (JFK), and Newark Liberty International Airport (EWR). That's right, over **three hundred thousand** flights, in these three airports alone. Impressive, right? So, how about we explore the data related to almost every single one of these flights? Because some flights have missing data (for example, cancelled flights have no departure and arrival time), I removed them from the data set to make our life a little easier. However, we still have over 300,000 flights to analyze. Have fun!

   Here is a description of the variables we have in the data set:

- *year*: the year of the flight.
- *month*: the month of the flight.
- *day*: the day of the flight.
- *dep_time*: departure time (format HHMM or HMM).
- *arr_time*: arrival time (format HHMM or HMM).
- *sched_dep_time*: scheduled departure time (format HHMM or HMM).
- *sched_arr_time*: scheduled arrival time (format HHMM or HMM).
- *dep_delay*: departure delay, in minutes. (Negative times represent early departures).
- *arr_delay*: arrival delay, in minutes. (Negative times represent early arrivals).
- *carrier*: the airline acronym (e.g., 'UL' is United Airlines).
- *flight*: flight number.
- *tailnum*: the plane tail number.
- *origin*: the origin airport.
- *dest*: the destination airport.
- *air_time*: the amount of time spent in the air, in minutes.
- *distance*: the distance between the airports, in miles.
- *dep_day_period*: the period of the day (morning, afternoon, night) a flight departed.

The data set is stored in the file *flight.csv* on Canvas. Download the data set and load the data set into R. Take a look at some of the variables by using the *head* function.

```
> head(flights)
```

(a) Waiting at the airport for the plane to take-off is quite boring. Investigate how the departure delays of the flights from New York is distributed by plotting the histogram. Describe the shape of the distribution. Are they doing a good job of keeping the flights on time? Briefly explain your answer. [4 marks]

(b) The histogram is very useful to tell us about the distribution, but taking a look at the numerical summary values is also super helpful. Given the shape of the histogram you obtained in the previous question, what would you expect the mean departure delay to compare with the median departure delay? Compute both the mean and the median. Are the results compatible with what you expected? Briefly explain. [3 marks]

(c) Similarly, obtain the IQR and standard deviation of departure delay. Is it what you expect? [2 marks]

(d) Quantiles can be a weird concept to grasp in a theoretical lecture, but they are incredibly useful in practice. [4 marks]

   i. obtain the 25%, 50%, and 95.149% quantiles and interpret the results.

   ii. if you were to catch a flight departing New York in 2013, what would be the probability that your flight wouldn't be more than 90 minutes late?

(e) What part of the day (i.e., morning, afternoon, or night) is the worst time concerning departure delays to take a flight? To answer this question, create an appropriate plot that displays the distribution of departure delays in each period of the day. What do you observe? (Optional: can you think of a potential reason for these results?) [3 marks]

(f) Let us investigate if the arrival delay is associated with the departure delay. Since we have too many data points (over 300,000 flights), the scatter plot can be a little too crowded and heavy on the computer. We will use an adaptation and create a two-dimensional bin (like the histogram has one-dimensional bins). See the image below:

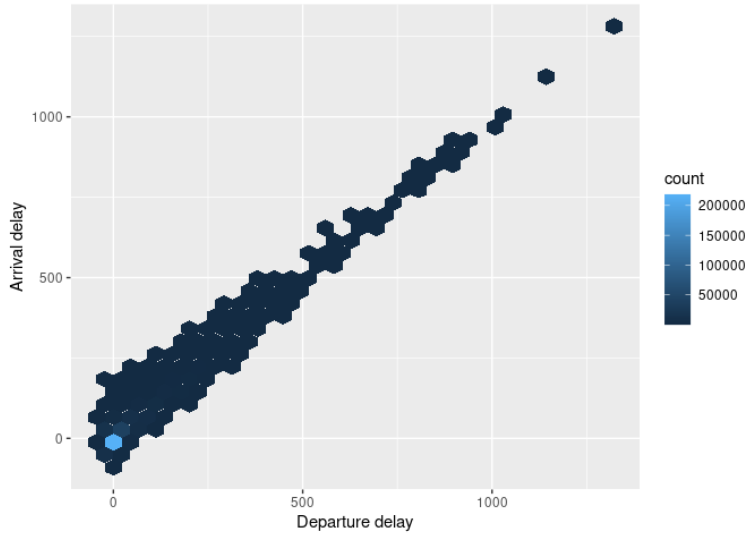

Figure 1:

Based on the plot above, would you say that arrival delay is associated with departure delay? If no, justify your answer. If yes, answer the following questions [3 marks]:

i. what would you say is the relationship's form (e.g., linear, quadratic, exponential, logarithmic)?

ii. how strong do you think the association is? Calculate any numeric summary/statistic to help you assess, if pertinent.

(g) Using the departure delay as explanatory variable and arrival delay as the response variable, calculate the coefficients of the regression line and write out your expression for the line of best fit. [2 marks]

(h) Provide an interpretation for the intercept and slope of the regression model obtained in the previous question. [3 marks]

(i) I fitted a linear model $predicted\ arr\_delay = b_0 + b_1 \times dep\_delay$. Below is the residual plot of my model. Assess the adequacy of fit and comment on the adequacy. [2 marks]

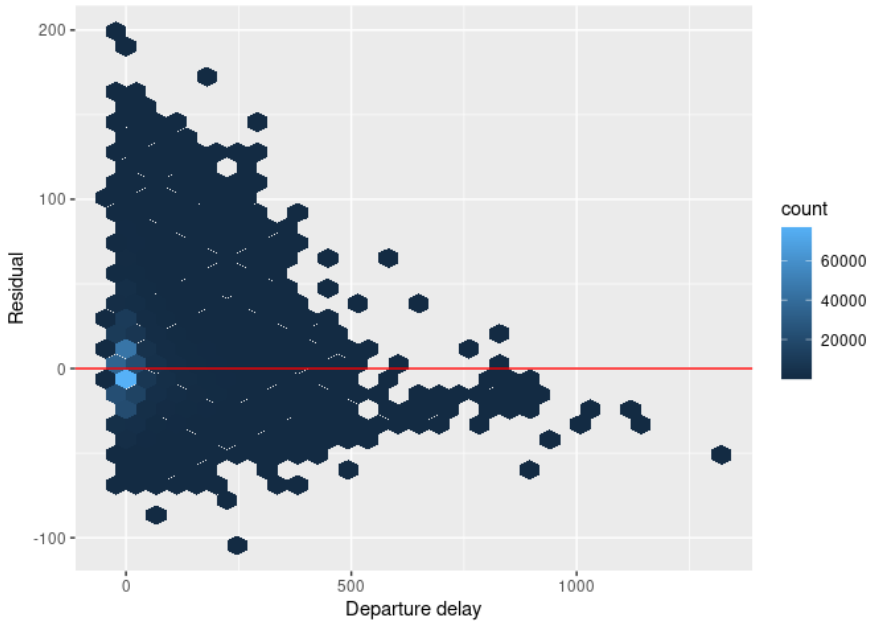

Figure 2:

(j) Your friend called you because she was bored at the airport waiting for her plane. You talked to her for a while. Suddenly, she mentioned that it was boarding time, after a 2 hours delay (finally!). She asked you if you thought she would arrive late at her destination and how late. Use the linear model from the previous question to answer her question. [2 marks]

(k) The table below shows the number of flights that were late departing from each of these 3 New York's airports.

origin	Less than 1 hour late	More than one hour late
Newark Liberty Int. Airport (EWR)	105803	11324
John F. Kennedy Int. Airport (JFK)	99976	9103
La Guardia Airport (LGA)	93250	7890

Based on this table, answer the following questions:

i. Your parents are coming to visit you from NY. Their flight will depart from EWR. What is the probability their flight will arrive more than one hour late? [1 mark]

ii. Your friend is coming to visit you from New York. You are waiting for her for over an hour past the scheduled arrival time. What is the probability that her flight departed from LGA? [1 mark]

iii. Do these data suggest an association between a flight from NY being over one hour late and the origin airport? Give statistical evidence to support your conclusion. [3 marks]